Title

Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models.

Journal

PLoS genetics, 12(3)

ISSN

1553-7390

Authors

Sul, Jae Hoon
Bilow, Michael
Yang, Wen-Yun
et al.

Publication Date

2016-03-01

DOI

10.1371/journal.pgen.1005849

Peer reviewed

PLOS | GENETICS

RESEARCH ARTICLE

Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models

Jae Hoon Sul[1☯], Michael Bilow[2☯], Wen-Yun Yang[2], Emrah Kostem[2], Nick Furlotte[2], Dan He[2], Eleazar Eskin[2,3]*

1 Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, California, United States of America, **2** Computer Science Department, University of California, Los Angeles, Los Angeles, California, United States of America, **3** Department of Human Genetics, University of California, Los Angeles, Los Angeles, California, United States of America

☯ These authors contributed equally to this work.
* eeskin@cs.ucla.edu

Citation: Sul JH, Bilow M, Yang W-Y, Kostem E, Furlotte N, He D, et al. (2016) Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. PLoS Genet 12(3): e1005849. doi:10.1371/journal.pgen.1005849

Editor: Nicholas J. Schork, University of California San Diego and The Scripps Research Institute, UNITED STATES

Received: March 26, 2015

Accepted: January 18, 2016

Published: March 4, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: JHS, WYY, EK, NF, DH and EE were supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, and 1320589, and National Institutes of Health grants K25-HL080079, U01- DA024417, P01-HL30568, PO1-HL28481, R01-GM083198, R01-MH101782 and R01-ES022282. EE is supported in part by the NIH BD2K award, U54EB020403. The funders had no

Abstract

Although genome-wide association studies (GWASs) have discovered numerous novel genetic variants associated with many complex traits and diseases, those genetic variants typically explain only a small fraction of phenotypic variance. Factors that account for phenotypic variance include environmental factors and gene-by-environment interactions (GEIs). Recently, several studies have conducted genome-wide gene-by-environment association analyses and demonstrated important roles of GEIs in complex traits. One of the main challenges in these association studies is to control effects of population structure that may cause spurious associations. Many studies have analyzed how population structure influences statistics of genetic variants and developed several statistical approaches to correct for population structure. However, the impact of population structure on GEI statistics in GWASs has not been extensively studied and nor have there been methods designed to correct for population structure on GEI statistics. In this paper, we show both analytically and empirically that population structure may cause spurious GEIs and use both simulation and two GWAS datasets to support our finding. We propose a statistical approach based on mixed models to account for population structure on GEI statistics. We find that our approach effectively controls population structure on statistics for GEIs as well as for genetic variants.

Author Summary

Although genome-wide association studies (GWASs) have discovered numerous novel genetic variants associated with many complex traits and diseases, those genetic variants typically explain only a small fraction of phenotypic variance. Factors that account for

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

phenotypic variance include environmental factors and gene-by-environment interactions (GEIs). Recently, several studies have conducted genome-wide gene-by-environment association analyses and demonstrated important roles of GEIs in complex traits. One of the main challenges in these association studies is to control effects of population structure that may cause spurious associations. In this paper, we show both analytically and empirically that population structure may cause spurious GEIs and use both simulation and two GWAS datasets to support our finding. We propose a statistical approach based on mixed models that can effectively correct for population structure when searching for GEIs.

Introduction

Over the past decade, genome-wide association studies (GWASs) have been a predominant approach to identify genetic variants involved in many complex traits and diseases.[1–3] While GWASs have discovered associations of many genetic variants, a large proportion of phenotypic variance for most traits is not explained by these variants.[4] Among several possible factors that explain this phenotypic variance such as effects of rare variants and epistasis, gene-by-environment interactions (GEIs) have drawn significant attention because of their important effect in many traits and diseases.[5–8] Discovering GEIs involved in diseases is of major interest in genetic research because they can provide insight into disease pathways, an understanding of the effect of environmental factors in disease, better risk prediction and personalized therapies. Similar to traditional GWASs that attempt to detect associations of genetic variants, researchers have recently performed gene-by-environment genome-wide association studies (GxE GWASs) to identify GEIs associated with diseases.[9–11]

One major difficulty in association studies is that population structure can easily confound the studies.[12] Association studies assume that individuals are unrelated, and if they are not, inflation of test statistics and possibly spurious associations may arise if genetic relatedness within individuals is imprecisely modeled. Several statistical approaches have been proposed to address this problem including genomic control [13], principal components analysis [14], and linear mixed models.[15] In particular, methods based on linear mixed models which incorporate pairwise relatedness between individuals has been shown to capture complex sample structure more effectively than other methods.[15, 16] It is important to note that all of these methods are designed to correct for population structure on statistics for genetic variants.

In contrast to numerous studies that have analyzed effect of population structure on association statistics in real GWAS datasets, few studies have investigated its effect on GEI statistics empirically. There have been, however, a few studies that evaluated bias caused by population structure on GEI statistics through simulations. Wang et al.[17] showed that population structure may have small effect on GEI statistics when genetic variants and environments have small correlations while Cheng and Lee [18] showed that it may introduce unacceptable bias to the estimation of GEIs in the presence of selection bias. Wang and Lee [19] also demonstrated that population structure may cause serious bias on estimated GEI effects in case-only studies. Recently, Dudbridge and Fletcher [20] showed that confounding due to population structure may cause dependence between gene and environment, and spurious GEIs can arise under this dependence. Although these studies provide useful information on theoretical impacts of population structure on GEI statistics, its influence in actual GxE GWASs has not been investigated comprehensively.

In this paper, we first show analytically that for the same reason that population structure causes spurious associations of genetic variants, it also causes spurious GEI associations based

on the polygenic model. We show that disregarding sample structure can easily inflate test statistics for GEIs, leading to false positives. We then simulate a GxE GWAS using the 1000 Genomes Project dataset.[21] This simulation demonstrates the impact of population structure on GEI statistics more accurately than previous simulations because it is based on actual genotype data that resemble traditional GWAS datasets whereas previous simulations are not. We show that test statistics for GEIs as well as those for genetic variants are inflated due to population structure.

In addition to the simulation, we utilize two GxE GWAS datasets to show that population structure may cause serious effects on GEIs. One dataset is an expression quantitative trait loci (eQTL) study of the human aortic endothelia cell collected by Romanoski et al.[22] and Erbilgin et al.[23] Gene expression was collected with and without a certain treatment, which corresponds to an environmental exposure. The other dataset is a GWAS dataset of inbred mouse strains termed Hybrid Mouse Diversity Panel (HMDP) that consists of 100 classical inbred and recombinant inbred strains.[24] We analyze their lipid phenotypes, and the environment exposure is a thioglycollate injection to recruit macrophages. Both datasets are ideal for evaluating effect of population structure on GEI statistics for following reasons. First, it is known that population structure exists in both datasets; individuals in the human eQTL dataset are from multiple ethnicities, and mouse strains in the HMDP dataset have very diverse genetic backgrounds. Second, both datasets have many quantitative phenotypes to test effect of GEIs; the human eQTL dataset has gene expression measured at more than 18,000 probes, and the HMDP dataset has more than 20 different quantitative phenotypes. This variety of phenotypes allows us to comprehensively determine the impact of population structure on GEI statistics.

We also propose a statistical approach based on a linear mixed model to correct for population structure on GEI statistics. We show that the traditional mixed model approach [15] that incorporates genetic relatedness between individuals only corrects for population structure on effects of genetic variants and does not correctly control inflation of test statistics for GEIs. To solve this problem, we consider two types of pairwise similarities between individuals. One is the traditional genetic similarity that causes a pair of individuals who are genetically similar to have correlated phenotypes, and this causes inflation of test statistics on genetic effects. The other type of similarity is that individuals who are related and have the same environment or exposure status have similar phenotypes, which causes spurious GEIs. We extend the linear mixed model to take into account both types of similarities and show that our approach effectively removes inflation of test statistics for both GEIs and genetic variants in our simulation and the two GxE GWAS datasets.

Results

Spurious GEIs due to population structure using 1000 Genomes simulation

We generate a simulated GxE GWAS using two populations (GBR and TSI) of 1000 Genomes Project dataset.[21] Each population has 1,000 individuals whose genotypes are generated using only common variants found in a standard SNP chip. In this simulation, we consider a dichotomous environmental exposure and two scenarios; (1) each population has the same number of exposed and unexposed individuals and (2) one population has more exposed individuals than the other population. We generate the genetic kinship matrix (K) from genotype data and the GxE kinship matrix (K^D) from K and the environmental exposure. Phenotypes are generated such that the genetic kinship (K) explains 40% of phenotypic variance while the GxE kinship (K^D) explains 20% (See Materials and Methods). There is no causal variant in the simulation, meaning that the genomic control inflation factor (λ_{GC}) should be close to one for

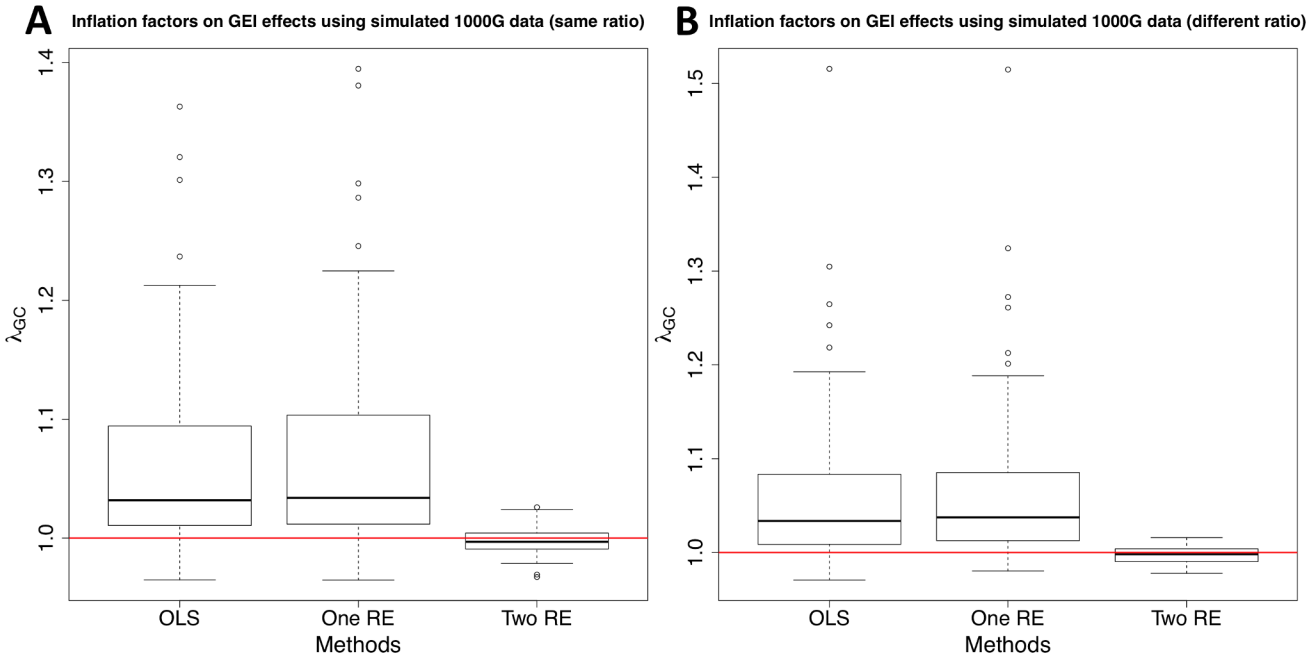

Fig 1. A distribution of inflation factors of GEI statistics on simulated 1000 Genomes data. We simulate genotype data using two populations (GBR and TSI), and genetic kinship (K) and GxE kinship (K^D) explain 40% and 20% of phenotypic variance, respectively. We generate 100 replicates of simulation, and measure inflation factors of three methods for each replicate; OLS, One RE, and Two RE. Y-axis is the inflation factor, and horizontal red line is drawn at $\lambda_{GC} = 1$. We assume a dichotomous environmental status where the two populations have the same number of exposed and unexposed samples (**A**) and where one population has more exposed samples than the other population (**B**).

doi:10.1371/journal.pgen.1005849.g001

both SNP and GEI statistics. We generate 100 replicates of simulation and measure inflation factors on SNP and GEI statistics of three different approaches. The first approach is one with no population structure correction on both SNP and GEI statistics ("OLS"), and another approach is a linear mixed model approach that incorporates the genetic kinship and accounts for population structure only on SNP statistics ("one RE"). The last approach is our proposed mixed model approach that uses both genetic and GxE kinship to correct for population structure on both SNP and GEI statistics ("two RE").

Fig 1 shows that population structure may cause spurious GEI associations because inflation factors on GEI statistics are on average greater than one. When the number of exposed and unexposed individuals is the same for both populations, the median λ_{GC} of the OLS approach is 1.032 and as high as $\lambda_{GC} = 1.363$. The results are similar when the ratio of exposed and unexposed individuals is different between the two populations. Population structure in the presence of GEIs may also cause inflation of SNP statistics, and S1 Fig shows that test statistics for SNPs are inflated. Also, λ_{GC} on SNP statistics tend to be higher than that on GEI statistics; the median inflation factor on SNP statistics is about 1.12. One of the reasons is that the genetic kinship (K) captures more phenotypic variance than the GxE kinship (K^D) does in this simulation. The result demonstrates that both SNP and GEI effects are susceptible to false associations due to population structure.

The result of the simulation also indicates that we need to incorporate both genetic and GxE kinship matrices into the linear mixed model to correct for population structure on SNP and GEI statistics. While the one RE approach that uses only genetic kinship reduces inflation of test statistics on SNPs (S1 Fig), it has almost the same or slightly worse inflation factors on GxE statistics than OLS (Fig 1). With our approach, λ_{GC} becomes very close to one; the median λ_{GC}

values on GEI statistics are 0.9969 and 0.9982 when the ratio of exposed and unexposed individuals between the two populations is the same and different, respectively. The maximum λ_{GC} values are also 1.026 and 1.0158, respectively. Interestingly, inflation factors on SNP statistics after applying our approach are even better than those after applying one RE; the median λ_{GC} with two RE is 0.9926 while that with one RE is about 1.02 when the ratio of exposed and unexposed individuals is the same. Hence, this shows that incorporating both kinship matrices also reduces inflation of test statistics on SNPs.

Human eQTL GxE GWAS results

To assess the influence of population structure on a real GxE GWAS, we first analyze the eQTL study of human aortic endothelial cell (HAEC).[22, 23] Erbilgin et al. measured gene expression levels of 147 individuals with and without the oxidized phospholipid species, oxidized 1-palmitoyl-2-arachidonoyl-snglycero-3-phosphatidylcholine (Ox-PAPC) treatment. In order to have independent samples, to perform a GxE GWAS, we randomly selected 74 samples where we only used the treated samples and 73 samples where we only used the untreated samples. Due to the normality assumption of the linear regression model, we filter out probes of gene expression that do not follow the normal distribution and choose 8,720 probes for our analysis (See Materials and Methods). We also perform the same quality control as in the original paper for the genotype data, and about 575,000 SNPs are included in our analysis. We compute λ_{GC} for each probe on SNP and GEI statistics of the three methods as in the previous simulation.

Fig 2 shows the distribution of inflation factors on GEI statistics with (Fig 2A) and without (Fig 2B) outliers. The results show that population structure indeed causes inflation of test

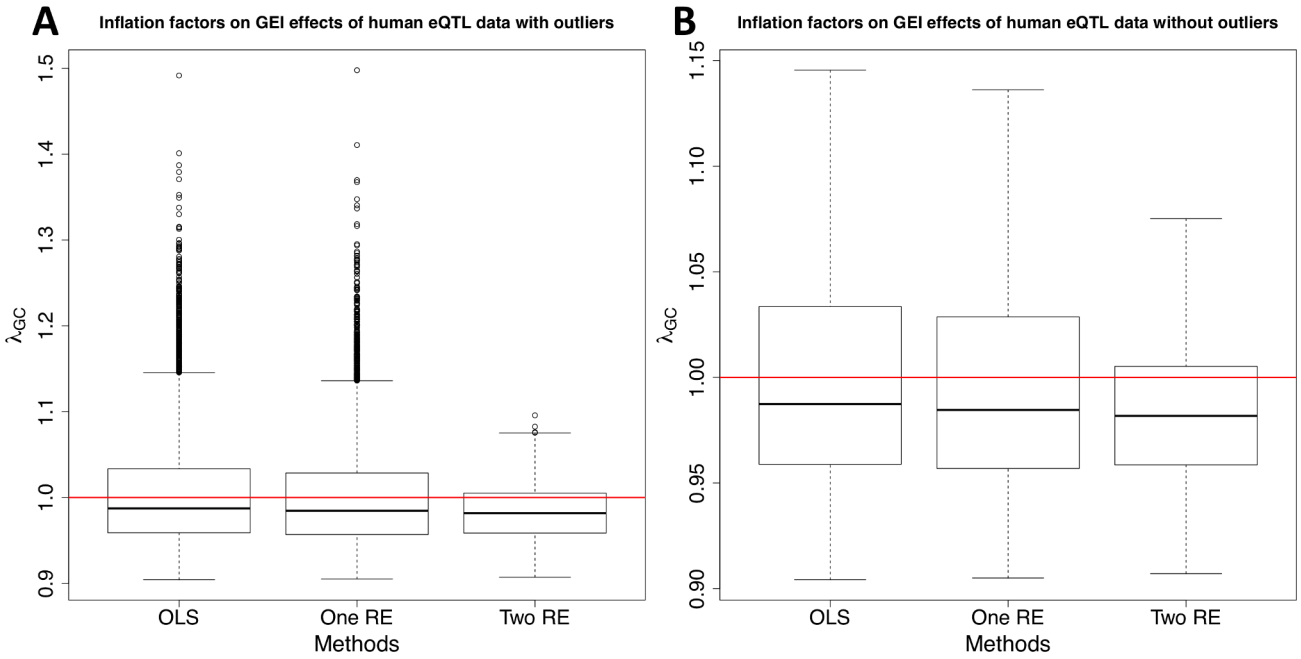

Fig 2. A distribution of inflation factors of GEI statistics on human eQTL GxE GWAS data. After filtering out probes whose expression values do not follow the normal distribution, 8,666 probes are tested for associations with about 500,000 SNPs. Gene expression of each individual was measured with and without the Ox-PAPC treatment, which corresponds to the environmental exposure. About a half of individuals were chosen to represent samples exposed to the environment, and the rest of individuals represent samples unexposed to the environment. We compute the inflation factor for each probe and for each of the three methods. Boxplots are drawn with outliers (**A**) and without outliers (**B**).

doi:10.1371/journal.pgen.1005849.g002

statistics for GEIs, and our method can effectively correct for population structure in a real GxE GWAS. Although all three approaches have very similar median inflation factors for GEI statistics (0.98), OLS and one RE approaches have many more probes whose λ_{GC} values are greater than one than our approach. There are 2,687 (31% of total probes) and 2,509 (29%) probes with $\lambda_{GC} > 1.02$ according to OLS and one RE approaches, respectively, and the maximum λ_{GC} values are 1.492 and 1.498, respectively. After applying our approach, there are only 950 probes (11% of total probes) with $\lambda_{GC} > 1.02$ and the maximum is 1.096. Fig 2B shows that even after removing outliers from the plot, our method has a narrower range of inflation factors than OLS and one RE approaches do. S2 Fig shows that our method also reduces inflation of test statistics on SNPs. Most of probes whose λ_{GC} values on SNP statistics are around or greater than 1.4 in the OLS approach have $\lambda_{GC} < 1.4$ after applying our method although the median λ_{GC} of our method is greater than one (1.0365).

We then compare the correlation between λ_{GC} and the variance of phenotype explained by the GxE kinship, denoted as $\hat{\sigma}^2_d$. We estimate variance components (σ^2_g, σ^2_d, σ^2_e in Eq (8)) using GCTA software [25], and we obtain the ratio of each variance component to the total phenotypic variance. We focus on only probes whose $\hat{\sigma}^2_d > 10\%$ because they are the probes in which the GxE kinship explains a certain amount of phenotypic variance. We find that about 24% (2,065) of probes have $\hat{\sigma}^2_d > 10\%$. Fig 3A shows that inflation factors of OLS on GEI statistics tend to increase as the variance of phenotype explained by the GxE kinship increases; r^2 between λ_{GC} and $\hat{\sigma}^2_d$ is 0.4631. This is expected because when σ^2_d is higher, GEI effects become more susceptible to false positives due to population structure. This is similar to higher λ_{GC} on SNP effects for phenotypes with higher σ^2_g. [15] r^2 between $\hat{\sigma}^2_d$ and λ_{GC} of the one RE approach (0.4620) is similar to that of the OLS approach (Fig 3B), meaning that it does not correct for population structure on GEI statistics. However, after applying our approach, r^2 becomes 0.0058 (Fig 3C). This means that even when the GxE kinship explains high phenotypic variance and hence population structure can easily confound GEI associations, our method can successfully correct for population structure.

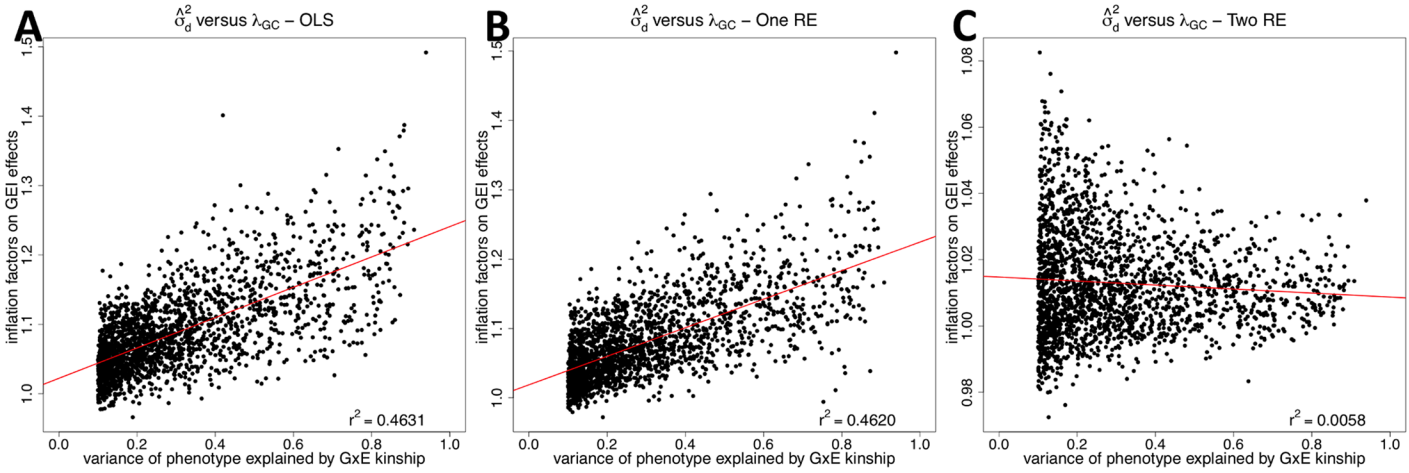

Fig 3. A correlation between the variance of phenotype explained by the GxE kinship matrix ($\hat{\sigma}^2_d$) and the inflation factor on GEI statistics (λ_{GC}) for human eQTL GxE GWAS data. The correlation is plotted for the three methods; OLS (**A**), One RE (**B**), and Two RE (**C**). Each dot is each probe, and x-axis is $\hat{\sigma}^2_d$ and y-axis is λ_{GC}. We estimate $\hat{\sigma}^2_d$ using the GCTA software, and only probes with $\hat{\sigma}^2_d > 10\%$ are shown in the plots. The red line is a regression line between λ_{GC} and $\hat{\sigma}^2_d$, and Pearson correlation coefficient is indicated on the top right of the plot.

doi:10.1371/journal.pgen.1005849.g003

HMDP GxE GWAS results

Next, we utilize the HMDP GxE GWAS dataset [24] that consists of many inbred mouse strains with very different genetic backgrounds. This diversity creates severe population structure, which was shown to easily cause spurious associations of SNP effects. [26] Hence, this dataset allows us to measure the impact of strong population structure on GEI statistics. We analyze 23 lipid phenotypes measured in more than 700 samples, and we test associations of about 74,000 SNPs after QC with these phenotypes. Macrophage recruitment was simulated in mice by injecting thioglycollate solution, which corresponds to environmental exposure in a GxE GWAS. The percentage of samples that received the injection varies between 30% to 42% for different phenotypes. We apply the three methods to each phenotype and measure the inflation factors on SNP and GEI statistics.

Fig 4A shows that population structure causes serious inflation of test statistics for GEIs; the median inflation factor of the OLS approach is 1.77. Inflation factors of the HMDP dataset are generally much greater than those of the human eQTL dataset, and this is expected because the HMDP dataset has much stronger population structure effect than the human eQTL dataset does. The results also show that λ_{GC} value becomes close to one and more stable after applying our approach. The median λ_{GC} of two RE is 1.092, and especially the maximum λ_{GC} is 1.19, which is much smaller than 6.27 of the OLS approach. Interestingly, the one RE approach has a worse distribution of inflation factors than the OLS approach as both median and maximum λ_{GC} values of one RE are much greater than those of OLS. This result is to a certain degree consistent with results of the previous 1000 Genome simulation; one RE tends to have higher λ_{GC} than OLS in the 1000 Genomes simulation. The one RE model performs similarly to the OLS model which demonstrates that traditional mixed model methods do not correct for GxE interactions. In fact, the one RE model performed slightly worse than the OLS model which is likely

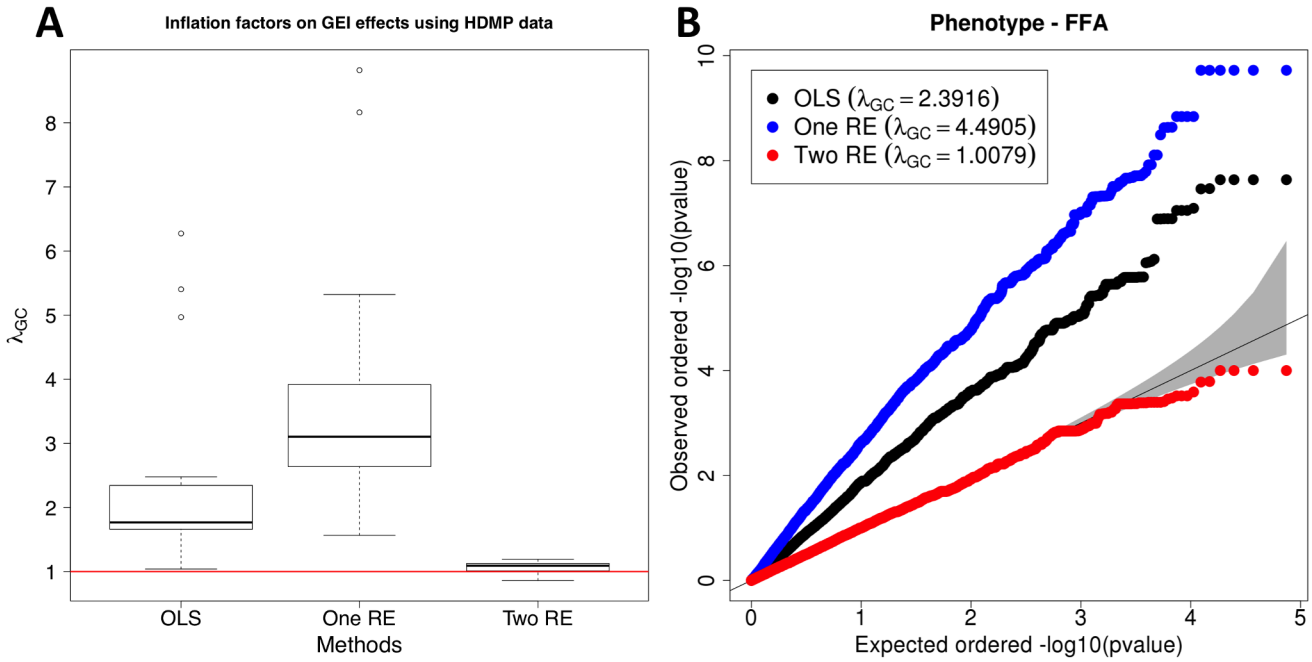

Fig 4. A distribution of inflation factors of GEI statistics on HMDP GxE GWAS data (A). HMDP consists of classical inbred strings and recombinant inbred strains. There are 23 lipid phenotypes, and about 74,000 SNPs are tested for associations. The environment is thioglycollate injection to recruit macrophages. We compute the inflation factor for each phenotype and for the three methods. (**B**) is a QQ plot of one of the phenotypes (free fatty acids, ffa), and it shows the distributions of p-values of GEI statistics for the three methods. Their inflation factors are indicated on the QQ plot.

doi:10.1371/journal.pgen.1005849.g004

Table 1. Variance of phenotype explained by the genetic kinship matrix ($\hat{\sigma}_g^2$), variance of phenotype explained by the GxE kinship matrix ($\hat{\sigma}_d^2$) and inflation factors for the three methods on GEI statistics for each phenotype of HMDP GxE GWAS data. Full name of each phenotype is discussed in Material and Methods section. GCTA software is utilized to estimate the phenotypic variance and its standard error for each phenotype.

Phenotype	Variance explained by K		Variance explained by K^D		Inflation factor on GEI statistics		
	$\hat{\sigma}_g^2$	SE	$\hat{\sigma}_d^2$	SE	OLS	One RE	Two RE
bw	61.10%	4.55%	8.23%	2.29%	1.6859	2.9499	1.1093
fat_mass	60.43%	4.93%	13.30%	3.13%	2.2928	4.1107	1.1432
ffa	41.12%	5.81%	11.30%	3.66%	2.3959	4.5087	1.0079
ffp	45.00%	6.76%	28.34%	5.58%	4.9688	8.1618	1.1137
ffp_percentage	41.03%	6.69%	26.90%	5.43%	5.4027	8.8199	1.1924
free_fluid	36.00%	5.01%	7.68%	2.62%	1.4134	2.2710	1.0297
gfp	58.07%	5.35%	16.57%	3.75%	2.3959	3.7234	1.1332
gfp_percentage	58.91%	5.18%	14.86%	3.47%	2.2350	3.5021	1.1419
glucose_lc	35.82%	5.85%	17.05%	4.17%	1.6833	3.1030	1.0276
glucose	34.48%	6.16%	19.43%	4.63%	1.6512	2.9215	0.8614
hdl	60.82%	4.39%	4.51%	1.67%	1.3387	3.4900	1.0396
ldl_and_vldl	40.63%	5.02%	6.21%	2.35%	1.6759	2.5118	0.9900
lean_mass	66.21%	4.16%	7.04%	1.99%	1.3440	3.2399	1.1883
mfp	48.07%	5.37%	12.05%	3.21%	1.9778	2.4283	1.1458
mfp_percentage	44.87%	5.39%	11.59%	3.19%	1.7668	2.2345	1.0916
nmr_bf_percentage	61.60%	4.88%	13.45%	3.13%	2.1546	4.1319	1.1058
nmr_total_mass	55.82%	5.10%	12.56%	3.08%	2.4778	3.6011	1.1017
rfp	53.52%	4.79%	5.67%	2.03%	1.7447	2.9424	1.0703
rfp_percentage	48.75%	4.93%	5.70%	2.12%	1.6711	2.8876	0.9809
spleen_wt	52.11%	4.93%	5.58%	2.14%	1.0402	1.5656	1.1195
tc	65.33%	4.05%	3.26%	1.34%	1.3364	2.4279	0.9781
tg	34.53%	5.72%	15.68%	4.00%	6.2736	5.3221	0.8774
uc	57.35%	4.99%	7.45%	2.59%	2.2306	2.7644	1.0137

doi:10.1371/journal.pgen.1005849.t001

because it is attempting to fix a statistical model which doesn't fit the data. Fig 4B is a QQ plot of one of the phenotypes, free fatty acids (ffa), and it shows that test statistics for GEIs from our method follow the expected distribution while those from the two other methods clearly have inflation of test statistics. S3 Fig shows λ_{GC} on SNP statistics, and the results are similar to those of the human eQTL dataset; both one RE and two RE approaches successfully removes inflation of test statistics on SNPs.

Table 1 lists the variance of phenotype explained by the genetic kinship matrix ($\hat{\sigma}_g^2$), one by the GxE kinship matrix ($\hat{\sigma}_d^2$) and inflation factors on GEI statistics for each phenotype. The genetic kinship matrix accounts for more phenotypic variance than the GxE kinship matrix for all phenotypes; the average $\hat{\sigma}_g^2$ is 50% while the average $\hat{\sigma}_d^2$ is 12%. However, for certain phenotypes, the GxE kinship explains more than 20% of phenotypic variance, and inflation factors on GEI statistics are greater for these phenotypes than for phenotypes with lower $\hat{\sigma}_d^2$. Fig 5 shows the correlation between λ_{GC} and $\hat{\sigma}_d^2$, and the OLS (Fig 5A) and one RE (Fig 5B) approaches have high correlations, which is similar to the results of the human eQTL dataset. However, our approach significantly reduces the correlation between $\hat{\sigma}_d^2$ and λ_{GC} (Fig 5C) meaning that our approach effectively removes effect of population structure.

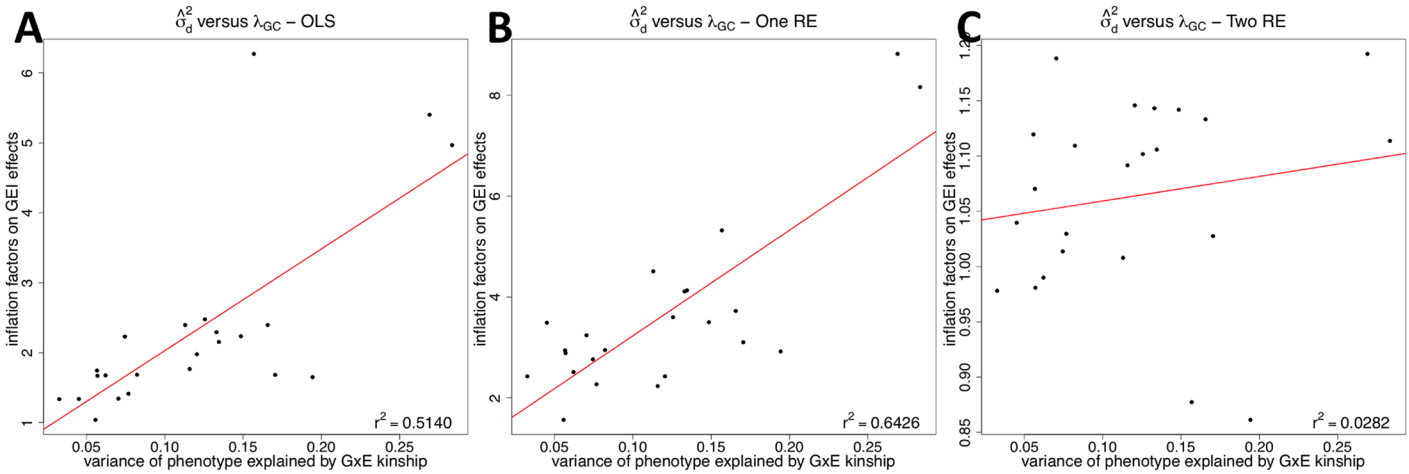

Fig 5. A correlation between the variance of phenotype explained by the GxE kinship matrix ($\hat{\sigma}^2_d$) and the inflation factor on GEI statistics (λ_{GC}) for HMDP GxE GWAS data. The correlation is plotted for the three methods; OLS (A), One RE (B), and Two RE (C). Each dot is each phenotype. The red line is a regression line between λ_{GC} and $\hat{\sigma}^2_d$, and Pearson correlation coefficient is indicated on the bottom right of the plot.

doi:10.1371/journal.pgen.1005849.g005

1-Random Effect Methods are often outperformed by OLS under significant popualtion structure

In Fig 4, we observe an anomalous result in HMDP GWAS data; the OLS results significantly *outperform* the 1RE method. This is somewhat surprising which computes a single variance component to account for the genetic heritability of the trait.

To determine why this is the case, we compare the amount of variance explained by K_G in the HMDP data with and without the use of a GxE study. To explore this, we simulated a population using the population structure derived from the HMDP data, as well as a population with a population with structure derived from our 1000G data. In this simulated data, we implanted several GxE associations and used GCTA to estimate the variance components of the model. These numbers are demonstrated in S1 Table. In this simulated data, we implanted several GxE associations and used GCTA to estimate the variance components of the model. We show that in the 1RE model, the variance components are substantially overestimates. We suspect that this is the cause of the higher observed inflation in the p-values compared to the OLS method.

Simulation framework of population structure on GEIs

To simulate a gene-by-environment (GxE) GWAS with population structure, we utilize HAPGEN2 software [27] to generate genotype data of two populations in 1000 Genomes Project [21]; GBR (British in England and Scotland) and TSI (Toscani in Italy). We use only common variants in chromosomes 11, 12, 13, and 14 whose minor allele frequency is greater than 5% in both populations. We also use variants present in the Illumina OmniExpress 730K genotyping chip to simulate a typical GWAS. The number of SNPs after this filtering is 99,612, and we generate 1,000 individuals for each population for a total of 2,000 individuals.

To generate phenotype values, we sample them from the following multivariate normal distribution.

$$\mathbf{y} \sim \mathcal{N}\left(0, \sigma^2_g \mathbf{K} + \sigma^2_d \mathbf{K}^{\mathrm{D}} + \sigma^2_e \mathbf{I}\right)$$

We create a genetic kinship matrix (or genetic relationship matrix) $\mathbf{K}$ using the GCTA software from the simulated genotype data. To create a GxE kinship matrix $\mathbf{K^D}$, we first need to assign an environmental exposure to each individual. We assume a dichotomous variable where a half of individuals (1,000) are exposed and the rest (1,000) are unexposed. When assigning an environmental exposure to each sample, we consider two possible cases. One is that each population has the same number of exposed and unexposed individuals. In other words, each population has exactly 500 exposed and 500 unexposed individuals. The other case is that one population has more exposed than unexposed individuals. For example, the ratio between exposed and unexposed in one population is 0.6 while it is 0.4 in other population. This is possible in actual GxE GWASs when individuals in one population are more easily exposed to the environment than those in other population. We vary this ratio from 0.54 to 0.6 in one population. We consider both cases in our simulation to determine how they influence results. Once we decide on the number of exposed and unexposed for each population, we randomly assign the environmental exposure to each individual and create the GxE kinship matrix. The phenotype values are generated such that the genetic kinship matrix and GxE kinship matrix explain 40% and 20% of phenotypic variance, respectively. In other words, σ_g^2 is 0.4, σ_d^2 is 0.2 and σ_e^2 is 0.4. We generate 100 replicates of this simulation.

Simulations of low σ_d^2

We also simulated our method at values of σ_d^2 (*i.e.* low GxE kinship) between 0 and 0.2, while holding the ratio between σ_g^2 and σ_e^2 constant; results are shown in S4 Fig. There appears to be an approximately linear relationship between the amount of inflation (λ_{GC}) and the size of σ_d^2. As less of the total variance is explained by the GxE kinship, the 2RE method begins to deflate p-values slightly, while both the 1RE and OLS methods improve. However, the over-correction of 2RE methods is small, and near $\sigma_d^2 = 0$, 1RE and 2RE methods perform similarly.

Bias and variance of $\hat{\sigma}_d^2$

Because the sample sizes in the HMDP and HAEC datasets are rather low for GWAS, we wanted to demonstrate that our method can accurately detect the amount of GxE kinship.

In order to do this, we performed simulations of populations with the same genetic kinship matrix as was computed from the HMDP dataset, and a phenotype distributed with an implanted GxE kinship component. The results, which demonstrate accurate estimation of this GxE component are in S6 Fig.

Principal components x environment studies

A alternate approach for correcting for population structure in gene-by-environmental interaction studies is to include principal components of the genetic relatedness matrix as covariates similar to the approach of EIGENSTRAT. We can extend such approaches to the scenario of gene-by-environment interactions by adding additional covariates of the form of the principal component times the environmental covariate (PCs × environment). We find that using PCs × environment does reduce inflation, but at a lesser extent than including a specifically GxE-based kinship matrix. This is consistent with comparisons of mixed models and principal components in traditional association studies [15]. We demonstrate these results in S7 Fig.

Effect of quantile normalization of phenotypes

Many of the phenotypes explored in the HMDP and HAEC datasets are close to normally-distributed but are not completely normal. We examine if this is a source of inflation by quantile

normalizing the data. In S5 Fig, we observe that quantile normalization of the phenotypes does improve the performance of 1RE and OLS methods, but only slightly.

Discussion

We demonstrated that population structure may cause spurious associations of gene-by-environment interactions. Using the same argument that population structure can inflate test statistics for genetic variants in the polygenic model, we were able to derive analytically that the same phenomenon may occur for GEI statistics. We then used the 1000 Genomes simulation and two GxE GWAS datasets to observe the impact of population structure on GEI statistics. When the severe population structure exists as in the mouse GxE GWAS dataset, we observed very high inflation factors for GEI statistics. When the influence of population structure is relatively moderate as in the 1000 Genome simulation and the human eQTL GxE GWAS dataset, we found that test statistics for GEIs are nonetheless inflated, which may cause spurious associations. While Wang et al.[17] showed through simulations that population structure may cause small biases to estimated GEI effects when there exists small correlation among environments and genetic variants, their results are not based on actual GxE GWAS datasets. Hence, our results that make use of current GxE GWASs may more accurately represent the impact of population structure on GEI statistics, and our results indicate that even moderate population structure may cause unacceptable inflation of test statistics for GEIs.

To correct for population structure on GEI statistics, we proposed a linear mixed model approach that includes two random effects to take into account two types of similarities between individuals. One is the genetic similarity, and the other is the similarity caused by both genetic and environment. By incorporating two kinship matrices corresponding to the two similarities into linear mixed models, we were able to correct for population structure on GEI statistics successfully. We showed that accounting for only the genetic similarity controls the inflation of test statistics for SNPs, but not for GEIs. This is important because GWASs typically include only the genetic kinship matrix to correct for population structure.[15] Sul and Eskin [16] proposed the idea of including two random effects in linear mixed models to account for two types of population structure; one caused by SNPs under selection and the other by rest of SNPs. As demonstrated in their and this papers, this approach is effective in removing inflation caused by two different types of population structure or confounding.

Recently, Zheng et al.[11] studied the roles of GEIs on type 2 diabetes (T2D) related traits and observed inflation of test statistics on GEIs. They collected information regarding to several dietary and lifestyle factors that may influence the T2D-related traits. These factors were considered as environmental exposure in their GxE analysis, and they measured the variance of T2D-related phenotype explained by GEIs for the different environmental factors. They also performed a GxE GWAS and observed that test statistics for GEIs were inflated for environmental factors that explained a significant amount of the phenotypic variance while they observed no inflation for factors that did not contribute to the phenotypic variance. One of the possible reasons for this inflation is population structure because they did not correct for population structure on GEI statistics. Their result is also consistent with our finding that the inflation factor is higher for a phenotype with higher $\hat{\sigma}^2_d$, the phenotypic variance explained by the GxE kinship. Hence, as more GxE GWASs are conducted to discover GEIs associated with traits, correcting for population structure will become important to reduce inflation of test statistics and to remove possible false positive associations.

The linear mixed model in our two RE approach is based on the GCTA GEI model [25], and we use GCTA software to estimate variance components. While previous GxE studies utilized GCTA software to estimate phenotypic variance explained by GEIs, to the best of our

knowledge, they did not attempt to measure effect of population on GEIs and to correct for population structure. Our approach is the first method to use the linear mixed model with two kinship matrices to correct for population structure on both SNP and GEI effects.

In this paper, we mainly focused on inflation of test statistics for measuring GEI effects of the three different approaches. We showed that only two RE approach achieves correct false positive rate while the two other methods do not. When comparing performance of different statistical tests, it is also important to compare power of tests in addition to measuring false positive rates. However, power comparison can only be made when all tests achieve the correct false positive rates. In our simulation and real data, it is not possible to compare power among the three different methods because OLS and one RE have incorrect false positive rate.

We note that in our results, we are including the marker that we are testing in the kinship matrix. Recently, several studies have pointed out that including the tested marker in the kinship matrix effectively includes the marker twice in the statistical model and this is what causes the inflation factor to be one even when there are many genetic effects throughout the genome. [28] Our results are orthogonal to these approaches and the recommendation of those studies should apply to testing for GEI as well. Nevertheless, in our experiments, we decided to include the tested marker in the kinship because this more easily exposes inflation of test statistics since we expect to observe an inflation factor of one if there is no inflation.

Materials and Methods

Spurious genetic associations due to population structure in the polygenic model

Before we discuss how population structure may cause spurious GEI associations, we first review how it influences associations of genetic variants because the two concepts are closely related. We assume that genetic effects are additive and there are M variants. Then, the standard genotype-phenotype model is

$$y_k = \mu + \sum_{i=1}^{M} \beta_i X_{ik} + \varepsilon_k \quad (1)$$

where y_k is individual k's phenotype value, μ is the mean of the phenotype, X_{ik} is the genotype of individual k at variant i, β_i is the effect of genetic variant i, and ε_k is the residual. The polygenic model assumes that there are many variants with small effects, which means that many β_i's are non-zero. The traditional association study considers each genetic variant individually and tests the effect of each genetic variant on the basis of the following model.

$$y_k = \mu + \beta_r X_{rk} + \eta_{\bar{r}k} \quad (2)$$

The goal of association studies is to identify the set of genetic variants with $\beta \neq 0$ since these are variants that putatively affect the phenotype. Note that we use different notations for the residual terms (ε_k in Eq (1) and $\eta_{\bar{r}k}$ in Eq (2)) to emphasize the difference between the two residuals. The residual in Eq (2) in relation to Eq (1) is exactly

$$\eta_{\bar{r}k} = \sum_{M:i\neq r} \beta_i X_{ik} + \varepsilon_k \quad (3)$$

According to Eq (3), people who are related would have similar residual terms ($\eta_{\bar{r}k}$) because they share the same genotypes (X_{ik}). This violates the assumption of the traditional linear regression model in Eq (2) that residuals are independent and hence causes bias in the estimation of β_r. Therefore, sample structure in GWAS datasets such as population structure or

cryptic relatedness may cause inflation of test statistics (β_r) for genetic variants.[12, 29] For clarity, we refer to the statistics testing for the effect of a genetic variant as SNP statistics to distinguish from statistics testing for the presence of GEI which we refer to as GEI statistics.

One approach to account for the sample structure is through the use of a linear mixed model.[15, 26, 30, 31] This approach introduces a random effect into the linear model in Eq (2) to account for the global genetic relatedness resulting in the following model

$$\mathbf{y} = \mu + \beta_r \mathbf{X_r} + \mathbf{u} + \mathbf{e} \tag{4}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ and $\mathbf{X_r} = [X_{r1}, X_{r2}, \ldots, X_{rn}]^T$ where n is the number of individuals. $\mathbf{u}$ is the random effect in the mixed model that captures effect of population structure, and $\mathrm{var}(\mathbf{u}) = \sigma_g^2\mathbf{K}$ and $\mathrm{var}(\mathbf{e}) = \sigma_e^2\mathbf{I}$ where $\mathbf{K}$ is an $n \times n$ kinship matrix and $\mathbf{I}$ is an identify matrix of size n. Then, the total variance of phenotype is given as $\mathrm{var}(\mathbf{y}) = \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}$. It has been shown that this linear mixed model approach that incorporates the pairwise genetic relatedness into the linear model effectively controls inflation of test statistics for genetic variants due to sample structure.[15, 26]

Spurious gene-by-environment interactions due to population structure

We extend the standard model to consider an environmental factor D. For simplicity, we assume it is a dichotomous variable. The exposure of an individual k to the environmental factor is denoted as D_k; $D_k = 0$ for the unexposed and $D_k = 1$ for the exposed. The model corresponding to Eq (1) is now

$$y_k = \mu + \sum_{i=1}^{M} \beta_i X_{ik} + \delta D_k + \sum_{j=1}^{M} \gamma_j D_k X_{jk} + \varepsilon_k \tag{5}$$

where δ represents the fixed effect of environmental factor D and each γ_j is the gene-by-environment interaction effect of variant j and environmental factor D. The goal of GxE association studies is to discover genetic variants or SNPs whose $\gamma_j \neq 0$ because they have effects on the phenotype in the presence of the environmental factor. While it appears from Eq (5) that interaction effects only affect the phenotype when $D_k = 1$, this is not always the case. For example, when $\beta_i = -\gamma_j$ and $D_k = 1$, a SNP does not influence the phenotype because SNP effects and interaction effects cancel each out. In this case, the SNP has effects on phenotype only for unexposed individuals ($D_k = 0$). Similar to the association study of genetic variants, we test the effect of each genetic variant and its GxE effect individually, which corresponds to fitting the following model.

$$y_k = \mu + \beta_r X_{rk} + \delta D_k + \gamma_k D_k X_{jk} + \tau_{\bar{r}k} \tag{6}$$

where $\tau_{\bar{r}k}$ is the residual. This residual is precisely

$$\tau_{\bar{r}k} = \sum_{M:i\neq r} \beta_i X_{ik} + \sum_{M:j\neq r} \gamma_j D_k X_{jk} + \varepsilon_k \tag{7}$$

These residuals ($\tau_{\bar{r}k}$) are not independent if individuals are related. For people who are genetically similar, they would have similar value for the first sum ($\Sigma_{M:\, i\neq r}\, \beta_i\, X_{ik}$), and people who are genetically similar and are in the same environment, they would have similar value for the second sum ($\Sigma_{M:\, j\neq r}\, \gamma_j\, D_k\, X_{jk}$). Hence, this equation shows that for the same reason population structure causes spurious genetic associations as shown in Eq (2), it may inflate test statistics of GEIs and cause false positive associations due to correlated residuals.

Linear mixed model to correct for population structure on GEIs

We extend the linear mixed model approach to correct for population structure on GEI statistics by introducing an additional random effect that captures the similarity of individuals due to GEI effects. Given the kinship matrix (K), we define the matrix K^D where each entry $K^D_{ij} = K_{ij}$ if $D_i = D_j$ and $K^D_{ij} = 0$ otherwise.[25] This matrix K^D describes how individuals are related both genetically and environmentally because a pair of individuals who are genetically related and share the same environment exposure have a non-zero kinship coefficient. We name K^D "GxE kinship" and K "genetic kinship" to distinguish two kinship matrices. We propose the linear mixed model that incorporates both kinship matrices as following

$$\mathbf{y} = \mu + \beta_r \mathbf{X_r} + \delta \mathbf{D} + \gamma_r \mathbf{D} \cdot \mathbf{X_r} + \mathbf{u} + \mathbf{v} + \mathbf{e} \tag{8}$$

where $\mathbf{D} = [D_1, D_2, \ldots, D_n]^T$ is a column vector of environmental exposures and $\mathbf{D} \cdot \mathbf{X_r}$ is the element-wise product. The random effect $\mathbf{v}$ accounts for the relatedness of individuals due to GEI effects and $\mathrm{var}(\mathbf{v}) = \sigma_d^2 \mathbf{K^D}$. The total variance of $\mathbf{y}$ is then given as $\mathrm{var}(\mathbf{y}) = \sigma_g^2 \mathbf{K} + \sigma_d^2 \mathbf{K^D} + \sigma_e^2 \mathbf{I}$. We call this approach "**two RE**" because it uses two random effects to correct for population structure on both SNP and GEI statistics.

We compare our approach to other approaches that do not consider effect of population structure on GEI statistics. One such approach is a simple linear regression without any random effect. We name this approach "**OLS**" from ordinary least squares, and it is defined as

$$\mathbf{y} = \mu + \beta_r \mathbf{X_r} + \delta \mathbf{D} + \gamma_r \mathbf{D} \cdot \mathbf{X_r} + \mathbf{e} \tag{9}$$

Note that this does not correct for population structure either on SNP statistics or on GEI statistics. Another approach is to correct for population structure only on SNP statistics by including one random effect that accounts for genetic relatedness. Its model is

$$\mathbf{y} = \mu + \beta_r \mathbf{X_r} + \delta \mathbf{D} + \gamma_r \mathbf{D} \cdot \mathbf{X_r} + \mathbf{u} + \mathbf{e} \tag{10}$$

This approach would account for the similarity due to genetic effects (the first sum in Eq (7)), but would not correct for the similarity due to GEI effects (the second sum in Eq (7)). This is because it is likely that values for β_i and γ_j are different for each variant, and the random effect $\mathbf{u}$ would not capture GEI effects which is the second sum in Eq (7). We name this approach "**one RE**" because it uses only one random effect.

P-values of all three approaches can be estimated using a standard F-test. Let $\Sigma = \mathbf{I}$ for OLS, $\Sigma = \hat{\sigma}_g^2 \mathbf{K} + \hat{\sigma}_e^2 \mathbf{I}$ for one RE, and $\Sigma = \hat{\sigma}_g^2 \mathbf{K} + \hat{\sigma}_d^2 \mathbf{K^D} + \hat{\sigma}_e^2 \mathbf{I}$ for two RE where $\hat{\sigma}_g^2, \hat{\sigma}_d^2, \hat{\sigma}_e^2$ are estimated variance components. We utilize GCTA software [25] to estimate these variance components ($\sigma_g^2, \sigma_d^2, \sigma_e^2$), and we estimate them for each phenotype once and apply them for all SNPs. This is the same as the EMMAX approach [15], and this approach markedly reduces the computational time while maintaining the similar power to that of an approach that estimates variance components for each SNP.

Let β include effects of all covariates in the linear regression model that includes a SNP effect (β_r) and a GEI effect (γ_r) and let $\mathbf{X}$ include all covariates including a SNP ($\mathbf{X_r}$) and a GEI ($\mathbf{D} \cdot \mathbf{X_r}$). Then, the estimated β is

$$\hat{\beta} = \left(\mathbf{X}' \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{y} \tag{11}$$

We perform a standard F-test for the null hypothesis $\beta_r = 0$ and $\gamma_r = 0$ to obtain p-values for SNP and GEI effects, respectively. We provide a software package that implements the two RE approach at http://genetics.cs.ucla.edu/pylmm/. Our approach can efficiently be applied to

standard GWAS datasets that contain thousands of individuals and hundreds of thousands of SNPs, similar to the linear mixed models for GWAS [15].

Exploring GEI of continuous environmental factors

It is very attractive to extend this model to the case of continuous covariates; however, it is not straightforward to create K^D from K. Under a binary enviornmental exposure, setting $K^D_{i,j} = K_{i,j} \times \delta_{D_i,D_j}$ where $\delta_{D_i,Dj}$ is 1 if the environmental exposures of indivdiuals i and j are the same and 0 otherwise is intuitive. For continuous covariates, we can extend this formalism to: $K^D_{i,j} = K_{i,j} \times f(D_i, D_j)$ for some function f. A good general guideline for a choice of function is to use $f = 1 - d(D_i, D_j)$ where d is a metric with range [0, 1]. Two of the most natural choices that satisfy the above recommendation are: $f(i,j) = 1 - |\frac{D_i - D_j}{R}|$, where R is the range of the environmental exposures, or $f(i,j) = 1 - |\mathbf{\Phi}\left(\frac{D_i - \mu_D}{\sigma_D}\right) - \mathbf{\Phi}\left(\frac{D_j - \mu_D}{\sigma_D}\right)|$, where μ_D and σ_D are the mean and standard deviation of the environmental exposures, and Φ is the standard normal cumulative distribution function. However, the best choice of f will depend on the scale of the environmental exposure.

Human eQTL GxE GWAS dataset

Erbilgin et al.[23] performed the expression quantitative trait loci (eQTL) study of human aortic endothelial cell (HAEC). They collected HAEC cultures from 147 unrelated heart transplant donors, and the oxidized phospholipid species, oxidized 1-palmitoyl-2-arachidonoyl-snglycero-3-phosphatidylcholine (Ox-PAPC) treatment was applied to the cells. It has been known that Ox-PAPC promotes vascular inflammation and regulates more than 1,000 transcripts in this cell type.[22, 32] Gene expression was collected both with and without the Ox-PAPC treatment. In order to have the two conditions have independent samples, we randomly chose a subset of 74 individuals where we only used the treated samples and a different 73 individuals where we only used the untreated samples, which represents two exposure statuses of environment.

The gene expression on 18,630 probes is collected using Affymetrix HT HG-U133A microarrays. The COMBAT software was utilized to correct for batch effects in the expression data [33]. Since the linear regression model assumes that expression values follow the normal distribution, we filtered out probes whose Shapiro-Wilk test p-values are less than 0.05. Additionally, we computed the number of outliers for each probe whose expression values are two standard deviations apart from the mean, and excluded probes containing five or more outliers (5% of total samples). These filters removed 9,910 probes, leaving 8,720 probes for the subsequent analysis. To verify that our results are not affected by the fact that the data still deviates from the normal distribution, we reanalyzed the data after performing quantile normalization both within each exposure group and over the entire dataset. In these additional experiments we observed equivalent results as shown in S5 Fig.

SNPs are genotyped using Affymetrix Genome-Wide Human SNP Array 6.0. We used the same QC filters as in the original paper [23]: MAF of 10%, HWE p-value of 10^{-4}, and genotype completeness of 5%, and 575,042 SNPs in autosomes are tested for associations. Erbilgin et al. performed a principal component analysis to identify population structure among the 147 individuals with 11 HapMap3 populations and found that there are groups of individuals with different ethnicities.

HMDP GxE GWAS dataset

Hybrid Mouse Diversity Panel (HDMP)[24] consists of 100 inbred strains including 29 classic inbred strains and three sets of recombinant inbred strains. The 23 lipid phenotypes (and their

abbreviations) that were analyzed are: body weight (bw), fat mass by NMR (fat_mass), free fatty acids (ffa), Femoral fat pad (ffp), Femoral fat pad/total bw (ffp_percentage), water weight by NMR (free_fluid), gonadal fat pad (gfp), gonadal fat pad/total bw (gfp_percentage), glucose at time for sac (glucose), glucose by lipid core (glucose_lc), HDL (hdl), LDL and VLDL (ldl_and_vldl), lean mass by NMR (lean_mass), mesenteric fat pad (mfp), mesenteric fat pad/total bw (mfp_percentage), Body fat percentage determined by NMR (nmr_bf_percentage), total mass by NMR (nmr_total_mass), retroperitoneal fat pad weight (rfp), retroperitoneal fat pad weight/total bw (rfp_percentage), spleen weight (spleen_wt), total cholesterol (tc), triglycerides (tg), unesterified cholesterol (uc). The number of samples from the 100 inbred strains that are phenotyped varies between 735 and 894 among these phenotypes. These strains are genotyped at more than 130,000 SNPs, and we applied following QC; genotype completeness of 98% for both SNPs and individuals and minor allele frequency threshold of 10%. After the QC, we have about 74,000 SNPs for performing association studies. The environment that we are interested in is thioglycollate injection to recruit macrophages. Macrophages play an important role in inflammatory component of many common diseases.[34] The percentage of exposed samples is between 30% and 42% depending on phenotypes. Note that individuals from the same strain can be both exposed and unexposed.

Supporting Information

S1 Fig. A distribution of inflation factors on SNP statistics on simulated 1000 Genomes data. Note that the scale is different from Fig 1. In **A**, the same number of exposed an unexposed individuals were generated in each simuation; in **B**, fewer exposed individuals were simulated compared to unexposed individuals.
(TIF)

S2 Fig. A distribution of inflation factors of SNP statistics on human eQTL GxE GWAS data. Note that the scale is different from Fig 2.
(TIF)

S3 Fig. A distribution of inflation factors of SNP statistics on HMDP GxE GWAS data. Note that the scale is different from Fig 4A.
(TIF)

S4 Fig. A distribution of the median inflation factors for SNP statistics on simulated 1000G GxE GWAS. The GxE variance σ_d^2 is taken from 0 to 0.2 while the purely genetic variance σ_g^2 and the purely random variance σ_e^2 are held constant.
(TIF)

S5 Fig. We explore the effect of quantile normalization between groups and within groups compared to the OLS estimate by itself. In **A**, we apply no quantile normalization; in **B**, the whole sample is first quantile normalized; in **C**, each environmental group is quantile-normalized separately. There is an apparent increase in performance of OLS and 1RE methods (their λ_{GC} becomes closer to 1) using quantile normalization, however, the performance is not as good as 2RE methods. There is little appreciable difference between the population-wide quantile normalization and within-group quantile normalization.
(TIFF)

S6 Fig. We demonstrate our method's ability to accurately measure implanted σ_d^2 through simulation. The bias is low, and the variance of the estimate appears to be somewhat proportional to the size of σ_d^2.
(TIF)

S7 Fig. We compare the performance of our 2-RE method with methods using a single random effect, regressing out the specified number of top principal components of the GxE kinship matrix from the output data.
(TIF)

S1 Table. Variance of phenotype explained by the genetic kinship matrix ($\hat{\sigma}_g^2$), when estimated in the absence of a GxE effect. The GxE heritability is highly significant for many phenotypes in Table 1, which this table references for the variances explained. GCTA software, using the environmental factor as a covariate, is utilized to estimate the phenotypic variance and its standard error for each phenotype.
(PDF)

Acknowledgments

We thank the reviewers for their constructive criticism in improving the text.

Author Contributions

Conceived and designed the experiments: JHS MB WYY EK NF DH EE. Performed the experiments: JHS MB WYY EK NF DH EE. Analyzed the data: JHS MB WYY EK NF DH EE. Contributed reagents/materials/analysis tools: JHS MB WYY EK NF DH EE. Wrote the paper: JHS MB WYY EK NF DH EE.

References

1. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–78. doi: 10.1038/nature05911 PMID: 17554300
2. Easton D. F., Pooley K. A., Dunning A. M., Pharoah P. D. P., Thompson D., Ballinger D. G., Struewing J. P., Morrison J., Field H., Luben R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447, 1087–93. doi: 10.1038/nature05887 PMID: 17529967
3. Schunkert H., König I. R., Kathiresan S., Reilly M. P., Assimes T. L., Holm H., Preuss M., Stewart A. F. R., Barbalic M., Gieger C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43, 333–8. doi: 10.1038/ng.784 PMID: 21378990
4. Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. A., Hunter D. J., McCarthy M. I., Ramos E. M., Cardon L. R., Chakravarti A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747–53. doi: 10.1038/nature08494 PMID: 19812666
5. Yang Q. and Khoury M. J. (1997). Evolving methods in genetic epidemiology. iii. gene-environment interaction in epidemiologic research. Epidemiol Rev 19, 33–43. doi: 10.1093/oxfordjournals.epirev.a017944 PMID: 9360900
6. Thomas D. (2010). Gene–environment-wide association studies: emerging approaches. Nat Rev Genet 11, 259–72. doi: 10.1038/nrg2764 PMID: 20212493
7. van IJzendoorn M. H., Bakermans-Kranenburg M. J., Belsky J., Beach S., Brody G., Dodge K. A., Greenberg M., Posner M., and Scott S. (2011). Gene-by-environment experiments: a new approach to finding the missing heritability. Nature Reviews Genetics 12, 881–881. doi: 10.1038/nrg2764-c1 PMID: 22094952
8. Hunter D. J. (2005). Gene–environment interactions in human diseases. Nature Reviews Genetics 6, 287–298. doi: 10.1038/nrg1578 PMID: 15803198
9. Hamza T. H., Chen H., Hill-Burns E. M., Rhodes S. L., Montimurro J., Kay D. M., Tenesa A., Kusel V. I., Sheehan P., Eaaswarkhanth M., et al. (2011). Genome-wide gene-environment study identifies glutamate receptor gene grin2a as a parkinson's disease modifier gene via interaction with coffee. PLoS Genet 7, e1002237. doi: 10.1371/journal.pgen.1002237 PMID: 21876681
10. Wei S., Wang L.-E. E., McHugh M. K., Han Y., Xiong M., Amos C. I., Spitz M. R., and Wei Q. W. (2012). Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. Carcinogenesis 33, 1531–7. doi: 10.1093/carcin/bgs188 PMID: 22637743

11. Zheng J.-S. S., Arnett D. K., Lee Y.-C. C., Shen J., Parnell L. D., Smith C. E., Richardson K., Li D., Borecki I. B., Ordovás J. M., et al. (2013). Genome-wide contribution of genotype by environment interaction to variation of diabetes-related traits. PLoS One 8, e77442. doi: 10.1371/journal.pone.0077442 PMID: 24204828

12. Helgason A., Yngvadóttir B., Hrafnkelsson B., Gulcher J., and Stefánsson K. (2005). An icelandic example of the impact of population structure on association studies. Nat Genet 37, 90–5. PMID: 15608637

13. Devlin B. and Roeder K. (1999). Genomic control for association studies. Biometrics 55, 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x PMID: 11315092

14. Price A. L., Patterson N. J., Plenge R. M., Weinblatt M. E., Shadick N. A., and Reich D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–9. doi: 10.1038/ng1847 PMID: 16862161

15. Kang H. M., Sul J. H., Service S. K., Zaitlen N. A., Kong S.-Y. Y., Freimer N. B., Sabatti C., and Eskin E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42, 348–54. doi: 10.1038/ng.548 PMID: 20208533

16. Sul J. H. and Eskin E. (2013). Mixed models can correct for population structure for genomic regions under selection. Nature Reviews Genetics. doi: 10.1038/nrg2813-c1 PMID: 23438871

17. Wang Y., Localio R., and Rebbeck T. R. (2006). Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions. Cancer Epidemiology Biomarkers & Prevention 15, 124–132. doi: 10.1158/1055-9965.EPI-05-0304

18. Cheng K. F. and Lee J. Y. (2012). Assessing the joint effect of population stratification and sample selection in studies of gene-gene (environment) interactions. BMC Genet 13, 5. doi: 10.1186/1471-2156-13-5 PMID: 22284162

19. Wang L.-Y. Y. and Lee W.-C. C. (2008). Population stratification bias in the case-only study for gene-environment interactions. Am J Epidemiol 168, 197–201. doi: 10.1093/aje/kwn130 PMID: 18497429

20. Dudbridge F. and Fletcher O. (2014). Gene-environment dependence creates spurious gene-environment interaction. The American Journal of Human Genetics. doi: 10.1016/j.ajhg.2014.07.014 PMID: 25152454

21. Abecasis G. R., Auton A., Brooks L. D., DePristo M. A., Durbin R. M., Handsaker R. E., Kang H. M., Marth G. T., and McVean G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65. doi: 10.1038/nature11632 PMID: 23128226

22. Romanoski C. E., Lee S., Kim M. J., Ingram-Drake L., Plaisier C. L., Yordanova R., Tilford C., Guan B., He A., Gargalovic P. S., et al. (2010). Systems genetics analysis of gene-by-environment interactions in human cells. Am J Hum Genet 86, 399–410. doi: 10.1016/j.ajhg.2010.02.002 PMID: 20170901

23. Erbilgin A., Civelek M., Romanoski C. E., Pan C., Hagopian R., Berliner J. A., and Lusis A. J. (2013). Identification of cad candidate genes in gwas loci and their expression in vascular cells. J Lipid Res 54, 1894–905. doi: 10.1194/jlr.M037085 PMID: 23667179

24. Bennett B. J., Farber C. R., Orozco L., Kang H. M., Ghazalpour A., Siemers N., Neubauer M., Neuhaus I., Yordanova R., Guan B., et al. (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res 20, 281–90. doi: 10.1101/gr.099234.109 PMID: 20054062

25. Yang J., Lee S. H., Goddard M. E., and Visscher P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011 PMID: 21167468

26. Kang H. M., Zaitlen N. A., Wade C. M., Kirby A., Heckerman D., Daly M. J., and Eskin E. (2008). Efficient control of population structure in model organism association mapping. Genetics 178, 1709–23. doi: 10.1534/genetics.107.080101 PMID: 18385116

27. Su Z., Marchini J., and Donnelly P. (2011). Hapgen2: simulation of multiple disease snps. Bioinformatics 27, 2304–5. doi: 10.1093/bioinformatics/btr341 PMID: 21653516

28. Listgarten J., Lippert C., Kadie C. M., Davidson R. I., Eskin E., and Heckerman D. (2012). Improved linear mixed models for genome-wide association studies. Nature methods 9, 525–526. doi: 10.1038/nmeth.2037 PMID: 22669648

29. Voight B. F. and Pritchard J. K. (2005). Confounding from cryptic relatedness in case-control association studies. PLoS Genet 1, e32. doi: 10.1371/journal.pgen.0010032 PMID: 16151517

30. Yu J., Pressoir G., Briggs W. H., Vroh Bi I., Yamasaki M., Doebley J. F., McMullen M. D., Gaut B. S., Nielsen D. M., Holland J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38, 203–8. doi: 10.1038/ng1702 PMID: 16380716

31. Zhao K., Aranzana M. J., Kim S., Lister C., Shindo C., Tang C., Toomajian C., Zheng H., Dean C., Marjoram P., et al. (2007). An arabidopsis example of association mapping in structured samples. PLoS Genet 3, e4. doi: 10.1371/journal.pgen.0030004 PMID: 17238287

32. Gargalovic P. S., Imura M., Zhang B., Gharavi N. M., Clark M. J., Pagnon J., Yang W.-P. P., He A., Truong A., Patel S., et al. (2006). Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci U S A 103, 12741–6. doi: 10.1073/pnas.0605457103 PMID: 16912112

33. Johnson W. E., Li C., and Rabinovic A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8, 118–27. doi: 10.1093/biostatistics/kxj037 PMID: 16632515

34. Orozco L. D., Bennett B. J., Farber C. R., Ghazalpour A., Pan C., Che N., Wen P., Qi H. X., Mutukulu A., Siemers N., et al. (2012). Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. Cell 151, 658–70. doi: 10.1016/j.cell.2012.08.043 PMID: 23101632